POLYFFUSION: A DIFFUSION MODEL FOR POLYPHONIC SCORE GENERATION WITH INTERNAL AND EXTERNAL CONTROLS

Lejun Min[1,2,4] **Junyan Jiang**[1,2] **Gus Xia**[1,2] **Jingwei Zhao**[3]
[1] Music X Lab, Computer Science Department, NYU Shanghai
[2] MBZUAI [3] Institute of Data Science, NUS
[4] Zhiyuan College, Shanghai Jiao Tong University
aik2mlj@gmail.com, {jj2731, gxia}@nyu.edu, jzhao@u.nus.edu

ABSTRACT

We propose Polyffusion, a diffusion model that generates polyphonic music scores by regarding music as image-like piano roll representations. The model is capable of controllable music generation with two paradigms: *internal* control and *external* control. Internal control refers to the process in which users pre-define a part of the music and then let the model infill the rest, similar to the task of masked music generation (or music inpainting). External control conditions the model with external yet related information, such as chord, texture, or other features, via the cross-attention mechanism. We show that by using internal and external controls, Polyffusion unifies a wide range of music creation tasks, including melody generation given accompaniment, accompaniment generation given melody, arbitrary music segment inpainting, and music arrangement given chords or textures. Experimental results show that our model significantly outperforms existing Transformer and sampling-based baselines, and using pre-trained disentangled representations as external conditions yields more effective controls. [1]

1. INTRODUCTION

Diffusion models [1, 2], as a new class of generative models, have been successful in generating high-quality samples of image data and beyond. They achieve state-of-the-art sample quality on a number of image generation benchmarks [3, 4], and also show strong results for the generation of various media such as audio [5, 6], video [7–9], and text [10, 11].

Symbolic music generation, a task very different from audio generation, has highly discrete outputs and is often described in terms of constraint optimization problems [12, 13]. Despite the improvement of deep music generative modeling [14, 15], symbolic music generation still suffers from the lack of controllability and consistency at different time scales [16]. In our study, we experiment with the idea of using diffusion models to approach controllable symbolic music generation.

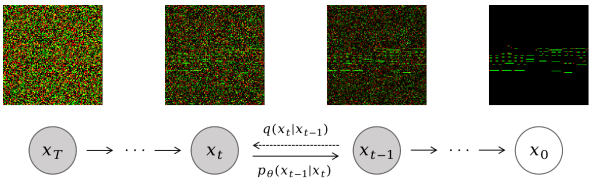


Figure 1: The forward and reverse process of the proposed diffusion model trained on piano roll representations. The red dot at the front of each note denotes its onset; the green bar following it denotes its sustain. Notice that the image axes are swapped for proper visualization.

Inspired by the high-quality and controllable image generation that diffusion models have achieved in computer vision, we devise an image-like piano roll format as the input, and used a UNet-based diffusion model to stepwise denoise a randomly sampled piano roll, as illustrated in Figure 1. We show in our experiments and demos that our design provides excellent generation results.

Besides unconditional generation, the model also accepts two categories of controls, namely internal control and external control:

- **Internal Control (Inpainting):** By masking out part of the given piano roll, we can specify the remaining area to be generated, thus implicitly conditioning the generation to fit in the masked part. We regard this strategy as a generalized music inpainting method.

- **External Control (Conditional Generation):** By adopting the cross-attention mechanism of Latent Diffusion [17], we can explicitly control the music generation on given external conditions such as chords and textures. They are first encoded into latent representations using pre-trained, disentangled variational autoencoders (VAEs), and then fed into the backbone UNet of the diffusion model to condition the denoising process. We show that the generated music complies with the given conditions well. We also add classifier-free guidance to control the variance of the generation.

[1] Demo page: https://polyffusion.github.io/. Code repository: https://github.com/aik2mlj/polyffusion

These controls of diffusion models enable us to unify a wide spectrum of creative music tasks that previously require separate modeling and training. In this paper, we showcase the following scenarios:

- **Melody generation given accompaniment** by generation with the accompaniment part being masked out.
- **Accompaniment generation given melody** by generation with the melody part being masked out.
- **Arbitrary music segment inpainting** by generation with any time segments being masked out.
- **General music arrangement given chords or textures** by conditioning on external chord or texture signals.

2. RELATED WORK

We review three realms of related work: 1) music inpainting, which is related to our internal control method, 2) conditioned music generation with external signals, which is related to our external control method, and 3) recent progress on diffusion-based modeling in the music domain.

2.1 Music Inpainting

Music inpainting is a controlled music generation task that regulates the generation with pre-defined musical contexts. We see various studies on polyphonic music inpainting. For example, DeepBach [18] develops a context-aware recurrent neural network (RNN) capable of inpainting missing notes for chorales in the style of Bach. Coconet [19] uses blocked Gibbs sampling to repeatedly rewrite a masked music score. Chang et al. [20] achieve variable-length music score inpainting. Music SketchNet [21] and MusIAC [22] introduce various controls to the inpainting task under VAE-based and Transformer-based framework respectively. Comparatively, diffusion models naturally possess the inpainting ability via masked generation [23], and there is no need to train or fine-tune a task-specific model for inpainting.

Though the current inpainting tasks mostly apply masks over a continuous period of time, the inpainted area, in theory, can be any note in the score (any area of a piano roll). In this study, we show that our image-like representation enables both part-wise and time-wise inpainting. The former refers to inpainting melody or accompaniment part given the other part, while the latter refers to infilling notes falling in arbitrary time segments.

2.2 Music Generation Conditioned on External Signals

External control signals are also one of the mainstream methods to control the music generation process. Common scenarios include generating music given chords [18, 24–26], lyrics [27], and other relevant features such as note density and voicing numbers [28].

Our study focuses on polyphonic score generation controlled by external chords and textures. In particular, the "control by texture" task has great practical value in both music arrangement and composition style transfer [29], while very few existing models could realize this function.

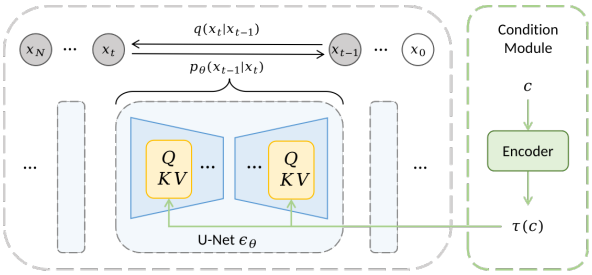


Figure 2: The model structure with an additional condition module for external control. Each UNet unit ϵ_θ applies one denoising step during the reverse process. External condition signals are encoded by pre-trained encoders and fed into the cross-attention layers, which are represented by the yellow squares in the UNet unit.

2.3 Diffusion Models for Music Generation

Recently, we have seen several attempts to introduce diffusion models to symbolic music tasks. Mittal et al. [30] generate monophonic music by training a diffusion model on the latent representations learned by MusicVAE [31]. Cheuk et al. [32] brings diffusion models to the music transcription task by adapting the piano roll format into the DiffWave [5] structure. It is relevant to our study as the model can also output piano rolls. However, the model focuses on transcription instead of generation by relying on a ground-truth spectrogram as its control. In general, for symbolic music generation, conditioning diffusion models on external controls is still an area to be explored.

3. METHODS

3.1 Data Representation

Our image-like piano roll representation is a 2-channel *binary* tensor $x \in \mathbb{R}^{2\times T\times P}$. The generation task targets 8-bar (32-beat) long music segments, with 1/4 beat as the time step, resulting in $T = 128$ time steps per sample. We use a MIDI pitch range 0...127, resulting in $P = 128$ pitch bins. Each entry $x(c, t, p)$ represents whether there is a note onset (for $c = 0$) or sustain (for $c = 1$) at time step t and MIDI pitch p.

3.2 Diffusion Model

Diffusion models [1, 2] are latent-variable models comprised of a forward (diffusion) process which gradually disrupts the structure of data x_0 and a reverse (denoising) process that learns to recover the original data x_0 from the noisy input. In our study, x_0 denotes the clean piano roll. The forward process iteratively adds Gaussian noise in N diffusion steps:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \quad (1)$$

$$q(x_{1:N}|x_0) = \prod_{t=1}^{N} q(x_t|x_{t-1}) \quad (2)$$

where $\beta_1, \beta_2, \ldots, \beta_N$ are a series of variance scheduling parameters. The reverse process requires the model to parameterize a Markov chain that iteratively reconstructs the piano roll x_0 from a corrupted input $x_N \sim \mathcal{N}(0, I)$.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)) \quad (3)$$

$$p_\theta(x_{0:N}) = p(x_N) \prod_{t=1}^{N} p_\theta(x_{t-1}|x_t) \quad (4)$$

During training, we optimize the model parameters ϵ_θ by minimizing the following target:

$$L(\theta) = \mathbb{E}_{x_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right] \quad (5)$$

where t is uniformly sampled from $[1, N]$ and $\epsilon \sim \mathcal{N}(0, I)$, $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. As shown in Figure 2, our unconditional model structure is based on [2], an image-oriented diffusion model using a 2-D UNet as its backbone ϵ_θ.

3.3 Internal Control (Inpainting)

Internal control refers to the use of the music notes themselves to regulate and influence the generation process, and we regard music inpainting as a means of internal control.

Specifically, we denote the given piano roll sample as s and the mask as m. At each step t during inference sampling, the fixed area of the image is diffused with the forward process $q(s_t|s) = \mathcal{N}(s_t; \sqrt{\bar{\alpha}_t}s, (1-\bar{\alpha}_t)I)$ and put together with the denoising sample s_{t-1}. Algorithm 1 [23] shows the detailed implementation of this inpainting process.

Algorithm 1 Inpainting Process

Input: inpainting mask m, original sample s, $x_N \sim \mathcal{N}(0, I)$

1: **for** $t = N, \ldots, 1$ **do**
2: $\quad \epsilon_1, \epsilon_2 \sim \mathcal{N}(0, I)$ if $t > 1$, else $\epsilon_1 = \epsilon_2 = 0$
3: $\quad y = \sqrt{\bar{\alpha}_t}s + \sqrt{1-\bar{\alpha}_t}\epsilon_1$ if $t > 1$, else s
4: $\quad x_{t-1} = \mu_\theta(x_t, t) + \sigma_\theta(x_t, t)\epsilon_2$
5: $\quad x_{t-1} = x_{t-1} \odot (1-m) + y \odot m$
6: **end for**
7: **return** x_0

3.4 External Control (Conditional Generation)

External control means using external signals to condition the generation process. We aim to incorporate a general strategy that does not place strong assumptions on the *format* of input control signals. To this end, we use the cross-attention mechanism [33] for conditional generation introduced by Latent Diffusion [17] since it is insensitive to the dimension of the condition signals. We also adopted the strategy used by Rombach et al. [17], which augments the backbone UNet structure with cross-attention layers that map condition signals into the UNet intermediate latent representations.

Formally, to preprocess the external musical signal c, we introduce a corresponding encoder τ that projects c to a latent representation $\tau(c)$. The encoder τ is pre-trained and fixed during diffusion model training. The cross-attention layers then map $\tau(c)$ to the intermediate layers of the UNet (as shown in Figure 2). The conditional training objective is

$$L_{\text{cond}}(\theta) := \mathbb{E}_{x_0,c,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, \tau(c)\right)\right\|^2\right] \quad (6)$$

We use classifier-free guidance (CFG) [34] to enable both conditioned and unconditioned generation by controlling the intensity of the condition signals during sampling. We refer readers to [34] and [35] for details on CFG.

4. CONTROLLABLE MUSIC GENERATION

In this section, we present four general musical applications our model empowers with internal and external controls: 1) melody generation given accompaniment, 2) accompaniment generation given melody, 3) arbitrary music segment inpainting, and 4) music arrangement given chords or textures. For each application, we provide non-cherry-picked generated samples as a case study. We also refer readers to our demo page for more examples.

4.1 Melody Generation Given Accompaniment

This task is achieved by internal control — to pre-define the accompaniment part and let the model infill the upper melody. Figure 3(a) shows an example of pop song melody generation given the accompaniment. We see that the melody is consistent with the underlying chords of the given accompaniment, and maintains an overall consistent rhythmic pattern, except for a 16th-note jump at the beginning of the 3rd bar.

4.2 Accompaniment Generation Given Melody

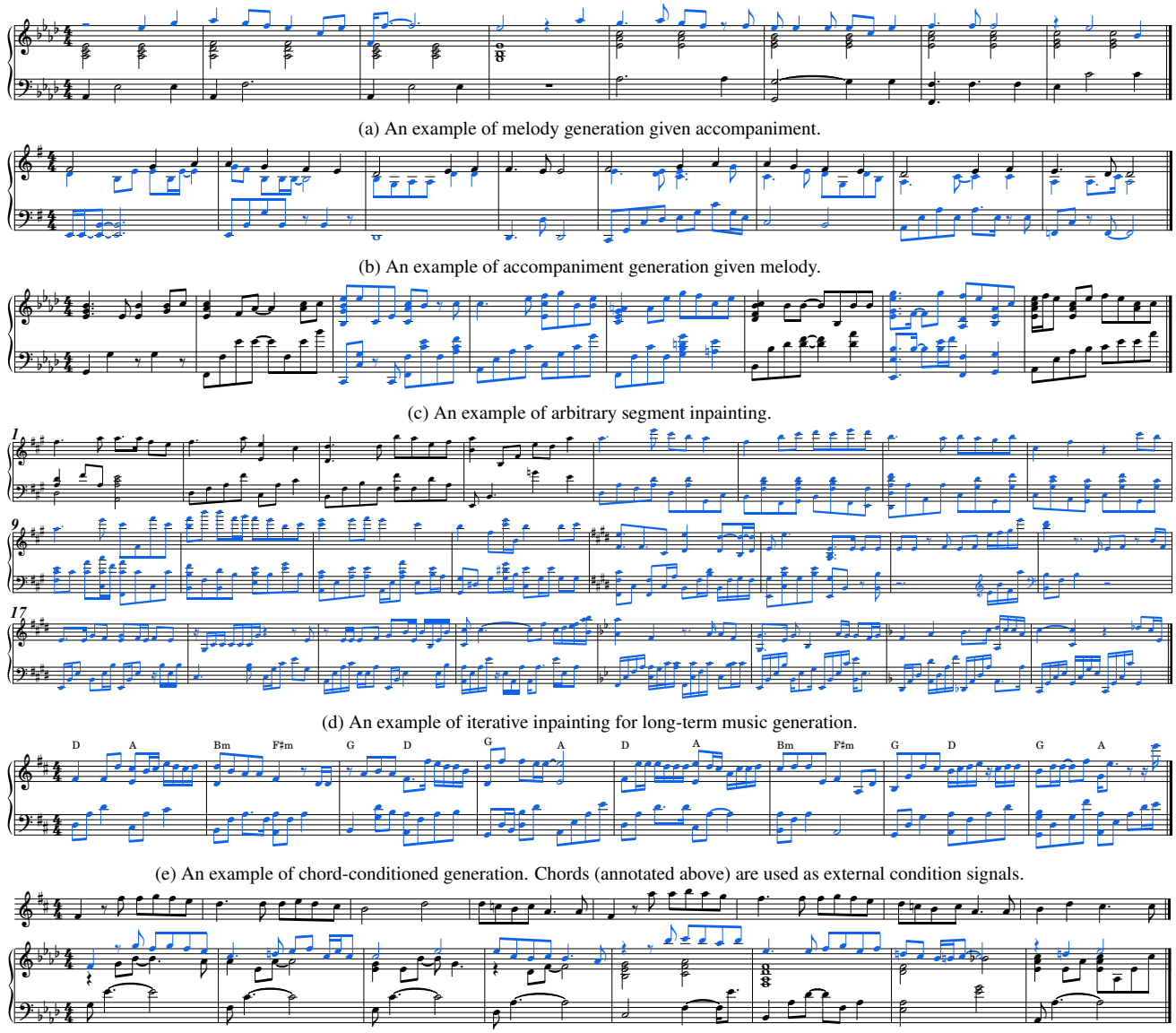
Similarly, given a lead melody, we can inpaint its corresponding lower accompaniment. Figure 3(b) shows an example, in which we see that the generated chord sequence suits the key (E minor) of the melody well, realized by a consistent arpeggio texture. The generated counter-melody also fills in the gaps between melody onsets well.

4.3 Arbitrary Music Segment Inpainting

The common scenario of music inpainting, also called music infilling [20], is to generate a music segment that fills in the gap between given past and future contexts. For our model, this task can be fulfilled by masking out the full pitch range of selected bars for inpainting.

Figure 3(c) shows an example of the inpainting process of the 3rd, 4th, 5th, and 7th bars, given the rest as fixed contexts. In the example, the model is capable of generating a full cadence connecting the 7th and the 8th bar, and also a nice applied chord in the non-diatonic progression Gm-Adim-B♭m connecting the 5th and the 6th bar.

We also extend the problem setup and let the diffusion model generate *long-term* music by iteratively inpainting



(a) An example of melody generation given accompaniment.

(b) An example of accompaniment generation given melody.

(c) An example of arbitrary segment inpainting.

(d) An example of iterative inpainting for long-term music generation.

(e) An example of chord-conditioned generation. Chords (annotated above) are used as external condition signals.

(f) An example of texture-conditioned melody generation. The texture of a given melody (the staff above) is used as external condition signals.

Figure 3: Generated samples in various tasks of controllable music generation. The generated parts are marked in blue. These examples have corresponding hearable demos on the demo page.

the future given the past. Figure 3(d) shows an example of a 24-bar generation based on a 4-bar prompt. The model generates 4 bars during each inference and finishes the process with five iterations. We see that the generated music contains a smooth chord progress, with a key modulation towards the end. The long-term textural structure is coherent, however lacking a consistent music theme.

4.4 Music Arrangement Given Chords or Textures

Inspired by the *chord-texture* disentanglement work [13, 29], we choose these two factors as the external condition signals for polyphonic generation. In our context, chords refer to the harmonic information, and textures refer to the rhythmic information. The latent chords and textures are encoded using pre-trained VAEs and cross-attended with the backbone UNet.

Beat-wise chords are first extracted by rule-based methods [36, 37], in which we adopted a 36-D chord representation consisting of a 12-D one-hot root encoding, a 12-D one-hot bass encoding and 12-D multi-hot chroma encoding. We then use a chord VAE [13] to extract a 512-D representation for each 8-bar chord sequence. For texture conditioning, we encode each 2-bar segment with the pre-trained texture encoder in [13] and then concatenate four encoded 256-D representations into a 1024-D vector as an 8-bar texture representation.

Figure 3(e) demonstrates an example of polyphonic music generation conditioned on chords. In the example, the accompaniment and the melody are mostly chord notes, with a certain degree of non-chord passing and neighboring tones that increase the interestingness of the song.

To show the complex combinations of conditions that the model can handle, we showcase a "texture-specified melody generation" for a given accompaniment segment as an example of the combination of internal and external

controls. As shown in Figure 3(f), We generate the melody part of a given accompaniment segment conditioned on the encoded texture representations of a given melody line. The result preserves a similar rhythmic pattern and fits the tonality of the new accompaniment.

5. EXPERIMENTS

5.1 Dataset and Training

We train our model using the POP909 dataset [38], a pop song dataset containing around 1K MIDI files. We only keep the pieces with 2/4 and 4/4 meters and cut them into 8-bar music segments with 1-bar hopping size, which results in 64K samples in total. The dataset is randomly split into the training set (90%) and validation set (10%) on a song level. The training samples are randomly transposed to all 12 keys for data augmentation.

The classifier-free guidance technique stated in Section 3.4 combines unconditional and conditional training. We adopt the implementation of DDPM and cross-attention layers in [39]. With 1K total diffusion steps, the model converges around 50 epochs (200K steps) on Adam Optimizer [40] with a constant learning rate 5e-5.

To turn the generated 2-channel piano roll representations into MIDI files, we round them to $\{0, 1\}$ and neglect notes without an onset. In practice, the generation process of 160 8-bar samples report zero invalid notes.

5.2 Evaluation

To validate the generation quality and control effectiveness of our model, we conducted both objective and subjective evaluations on 5 tasks: (1) unconditional generation, (2) accompaniment generation, (3) segment inpainting, (4) chord-conditioned generation, and (5) texture-conditioned generation. Tasks 2-3 focus on the evaluation of internal controls, and tasks 4-5 focus on external controls. Table 1 summarizes the evaluation method for each task.

5.2.1 Evaluation Metrics

Objective metrics: To objectively measure the music quality for all 5 tasks, we use the averaging overlapped area of pitch distribution ($\mathcal{D}_{\mathrm{P}}$) and duration distribution ($\mathcal{D}_{\mathrm{D}}$) from [41], which measure the distribution similarity of pitch and duration between the generated samples and ground truth. Additionally, we introduce *chord distance* (CD) [41] and *onset distance* (OD) to evaluate the efficacy of external control. These metrics measure the ℓ_2 distance of chord (for task 4) and onset distribution (for task 5) between the generated samples and the chord/texture condition.

Subjective metrics: Subjective metrics include *creativity* (C), *naturalness* (N), and *musicality* (M), which provide a perceptive evaluation complementing the objective musical quality metrics. To demonstrate the efficacy of internal control, we pick accompaniment generation as an example and add a *fitness* (F) metric to evaluate how well the generated parts fit in with the given melody.

5.2.2 Baseline models

We use two types of models as our baselines:

Transformer models: As suggested in the polyphonic representation disentanglement study [13], applying a Transformer on disentangled latent codes yields better results than raw token predictions. Following [13], we train a Transformer to predict the chord and texture representations from melody representations. For unconditional generation (task 1), we sample the latent spaces of the first 2-bar melody and then predict its accompaniment and the following content. For accompaniment generation (task 2) and external conditioning (tasks 4-5), the melody (task 2), chord (task 4), or texture (task 5) latent representation is directly encoded as the condition for the Transformer. We adopt the XLNet-based model proposed in [20] for the music segment inpainting task (task 3).

Sampling-based models: We adopt the VAE-based disentanglement model in [13] and generate music segments by sampling the latent spaces. For unconditional generation (task 1), we sample from the chord and texture latent spaces of the first and the last 2 bars, then linearly interpolate the middle latent codes to form a coherent 8-bar segment. For inpainting (task 3), we also use linear interpolation on latent codes to infill the missing bars. For external conditioning (tasks 4-5), the chord (task 4) or texture (task 5) latent component is directly encoded from the given condition.

5.3 Comparative Results

We calculate the average of each objective metric on 160 generated samples for each task. As shown in Table 2, Polyffusion and its variations achieve the highest objective scores in tasks 1-4. For controllability, our model yields competitive results on segment inpainting and chord-conditioned generation. For the texture-conditioned generation task, our model does not perform as well as the baseline but is capable of preserving the general musical texture, since the baseline model is explicitly trained on texture reconstruction targets, while the texture condition of our model only serves as a hint for the generation.

We also show the effectiveness of classifier-free guidance in Table 2. With a guidance scale of 5, the model (Polyffusion-S5) shows improved controllability on both chord conditioning and texture conditioning. Notably, a large guidance scale for chord conditions negatively impacts the $\mathcal{D}_{\mathrm{D}}$ metric. We speculate that this is because notes regular in length provide clearer chord context, which can be noticed in the guidance demos.

For subjective evaluation, we invite participants to rate the generation quality via a double-blind online survey. Our survey consists of 4 groups of samples of unconditional generation and accompaniment generation, respectively. Each group contains a ground-truth piece, generated samples by Polyffusion and all baselines with random orders. 36 participants completed our survey. Each participant rated 4 random groups on average based on a 5-point scale. The evaluation results are shown in Figure 4 and 5. The height of each bar represents the mean rating,

	(1) Uncond. Gen.	(2) Acc. Gen.	(3) Seg. Inp.	(4) Chord Cond.	(5) Texture Cond.
Objective Metrics	$\mathcal{D}_{\mathrm{P}}$, $\mathcal{D}_{\mathrm{D}}$	$\mathcal{D}_{\mathrm{P}}$, $\mathcal{D}_{\mathrm{D}}$	$\mathcal{D}_{\mathrm{P}}$, $\mathcal{D}_{\mathrm{D}}$	$\mathcal{D}_{\mathrm{P}}$, $\mathcal{D}_{\mathrm{D}}$, CD	$\mathcal{D}_{\mathrm{P}}$, $\mathcal{D}_{\mathrm{D}}$, OD
Subjective Metrics	C, N, M	C, N, M, F	N/A	N/A	N/A
Generative Length	8 bars	8 bars	4 bars	8 bars	8 bars
Transformer Baselines	Wang	Wang	Chang	Wang	Wang
Sampling Baselines	Wang	N/A	Wang	Wang	Wang

Table 1: Specifications of the evaluation tasks and the baseline models. C, N, M, F in subjective metrics mean creativity, naturalness, musicality, and fitness respectively. *Wang* refers to the Transformer models (for Transformer baselines) and VAE-based models (for sampling baselines) in [13]; *Chang* refers to the XLNet-based model in [20].

	Uncond. Gen.		Acc. Gen.		Seg. Inp.		Chord Cond.			Texture Cond.		
	$\mathcal{D}_{\mathrm{P}}\uparrow$	$\mathcal{D}_{\mathrm{D}}\uparrow$	$\mathcal{D}_{\mathrm{P}}\uparrow$	$\mathcal{D}_{\mathrm{D}}\uparrow$	$\mathcal{D}_{\mathrm{P}}\uparrow$	$\mathcal{D}_{\mathrm{D}}\uparrow$	$\mathcal{D}_{\mathrm{P}}\uparrow$	$\mathcal{D}_{\mathrm{D}}\uparrow$	CD $\downarrow$	$\mathcal{D}_{\mathrm{P}}\uparrow$	$\mathcal{D}_{\mathrm{D}}\uparrow$	OD $\downarrow$
Polyffusion	**0.89**	**0.93**	**0.89**	**0.96**	**0.90**	**0.93**	0.90	**0.96**	0.75	0.88	**0.98**	1.85
Polyffusion-S5	0.89	0.93	0.89	0.96	0.90	0.93	**0.92**	0.81	**0.51**	0.87	0.97	1.75
Polyffusion-A	0.89	0.93	0.89	0.96	0.90	0.93	0.90	0.94	0.79	**0.95**	**0.98**	4.37
Transformer	0.78	0.84	0.88	0.89	**0.90**	0.83	0.87	0.88	0.56	0.84	0.93	**0.13**
Sampling	0.86	0.90	N/A	N/A	0.89	0.91	0.86	0.90	0.70	0.91	0.93	0.20

Table 2: The objective evaluation and ablation study results. The statistics of generation, accompaniment generation and segment inpainting are identical for three Polyffusion models (hence gray-out for the latter two models) since they share the same internal control method.

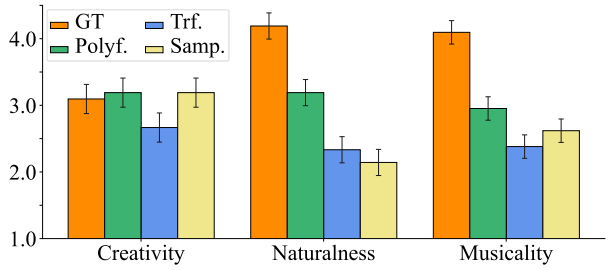


Figure 4: Subjective evaluation for unconditional generation.

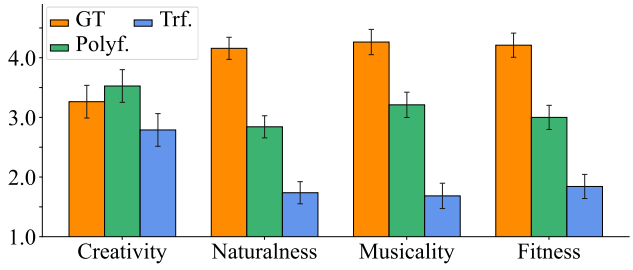


Figure 5: Subjective evaluation for accompaniment generation.

and the error bars are MSEs computed by within-subject ANOVA [42]. We report a significantly better performance (p-value < 0.05) of Polyffusion than baseline models in *naturalness* and *musicality* for both tasks and in *fitness* for accompaniment generation. Interestingly, Polyffusion even outperforms the ground truth on the *creativity* metric.

5.4 Ablation Study

We perform an ablation test on the use of VAE encoders for condition signals. For both chord conditioning and texture conditioning, we remove the corresponding pre-trained encoders. The ablated model of chord conditioning uses concatenated 36-D chord vectors as the condition signals. The ablated model of texture conditioning uses a modified piano roll representation [13]. Both models are trained with the same settings as the proposed model. Table 2 shows that the ablated models (Polyffusion-A) perform worse than the proposed models on the controllability metrics (CD & OD), showing the advantage of using disentangled latent representations as condition signals for diffusion models.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose a diffusion model for polyphonic symbolic music generation. We show that an image-like piano roll representation is effective for modeling the musical context for a high-quality score generation. We specify two methods for controllable generation: internal control via masked generation, and external control via conditioning using cross-attention. Experiments show that our method achieves higher quality and controllability compared to the Transformer and sampling-based baselines on both internal and external control tasks.

We regard the diffusion framework as a prospective direction for future work on controllable music generation, since it achieves fine-grained controls over high-quality generation and enables a wide spectrum of arrangement applications. Currently, our generation is limited to quantized music scores without performance features. We plan to extend this methodology to expressive performance modeling. Several new controls can also be introduced to facilitate human-AI co-creation of symbolic music, e.g., hierarchical structure controls (e.g., music segment labels) and multimodal controls (e.g., text descriptions).

7. REFERENCES

[1] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[3] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[4] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation." *J. Mach. Learn. Res.*, vol. 23, pp. 47–1, 2022.

[5] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[6] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," *arXiv preprint arXiv:2009.00713*, 2020.

[7] R. Yang, P. Srivastava, and S. Mandt, "Diffusion probabilistic modeling for video generation," *arXiv preprint arXiv:2203.09481*, 2022.

[8] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *arXiv preprint arXiv:2204.03458*, 2022.

[9] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood, "Flexible diffusion modeling of long videos," *arXiv preprint arXiv:2205.11495*, 2022.

[10] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," *arXiv preprint arXiv:2205.14217*, 2022.

[11] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, "Diffuseq: Sequence to sequence text generation with diffusion models," *arXiv preprint arXiv:2210.08933*, 2022.

[12] F. Pachet and P. Roy, "Musical harmonization with constraints: A survey," *Constraints*, vol. 6, no. 1, pp. 7–19, 2001.

[13] Z. Wang, D. Wang, Y. Zhang, and G. Xia, "Learning interpretable representation for controllable polyphonic music generation," *arXiv preprint arXiv:2008.07122*, 2020.

[14] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[15] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[16] J.-P. Briot and F. Pachet, "Deep learning for music generation: challenges and directions," *Neural Computing and Applications*, vol. 32, no. 4, pp. 981–993, 2020.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[18] G. Hadjeres, F. Pachet, and F. Nielsen, "Deepbach: a steerable model for bach chorales generation," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1362–1371.

[19] C.-Z. A. Huang, T. Cooijmans, A. Roberts, A. Courville, and D. Eck, "Counterpoint by convolution," *arXiv preprint arXiv:1903.07227*, 2019.

[20] C.-J. Chang, C.-Y. Lee, and Y.-H. Yang, "Variable-length music score infilling via xlnet and musically specialized positional encoding," *arXiv preprint arXiv:2108.05064*, 2021.

[21] K. Chen, C.-i. Wang, T. Berg-Kirkpatrick, and S. Dubnov, "Music sketchnet: Controllable music generation via factorized representations of pitch and rhythm," *arXiv preprint arXiv:2008.01291*, 2020.

[22] R. Guo, I. Simpson, C. Kiefer, T. Magnusson, and D. Herremans, "Musiac: An extensible generative framework for music infilling applications with multi-level control," in *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*. Springer, 2022, pp. 341–356.

[23] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 461–11 471.

[24] I. Simon, D. Morris, and S. Basu, "Mysong: automatic accompaniment generation for vocal melodies," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2008, pp. 725–734.

[25] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Generating music with rhythm and harmony," *arXiv preprint arXiv:2002.00212*, 2020.

[26] C. Donahue, H. H. Mao, Y. E. Li, G. W. Cottrell, and J. McAuley, "Lakhnes: Improving multi-instrumental music generation with cross-domain pre-training," *arXiv preprint arXiv:1907.04868*, 2019.

[27] Z. Ju, P. Lu, X. Tan, R. Wang, C. Zhang, S. Wu, K. Zhang, X. Li, T. Qin, and T.-Y. Liu, "Telemelody: Lyric-to-melody generation with a template-based two-stage method," *arXiv preprint arXiv:2109.09617*, 2021.

[28] J. Zhao and G. Xia, "Accomontage: Accompaniment arrangement via phrase selection and style transfer," *arXiv preprint arXiv:2108.11213*, 2021.

[29] S. Dai, Z. Zhang, and G. G. Xia, "Music style transfer: A position paper," *arXiv preprint arXiv:1803.06841*, 2018.

[30] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, "Symbolic music generation with diffusion models," *arXiv preprint arXiv:2103.16091*, 2021.

[31] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," in *International conference on machine learning*. PMLR, 2018, pp. 4364–4373.

[32] K. W. Cheuk, R. Sawata, T. Uesaka, N. Murata, N. Takahashi, S. Takahashi, D. Herremans, and Y. Mitsufuji, "Diffroll: Diffusion-based generative music transcription with unsupervised pretraining capability," *arXiv preprint arXiv:2210.05148*, 2022.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[34] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[35] S. Dieleman, "Guidance: a cheat code for diffusion models," 2022. [Online]. Available: https://benanne.github.io/2022/05/26/guidance.html

[36] B. Pardo and W. P. Birmingham, "Algorithms for chordal analysis," *Computer Music Journal*, vol. 26, no. 2, pp. 27–49, 2002.

[37] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "Mir_eval: A transparent implementation of common mir metrics." in *Proceedings of the 15th International Conference on Music Information Retrieval*, 2014.

[38] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, X. Gu, and G. Xia, "Pop909: A pop-song dataset for music arrangement generation," *arXiv preprint arXiv:2008.07142*, 2020.

[39] N. W. Varuna Jayasiri, "labml.ai annotated paper implementations," 2020. [Online]. Available: https://nn.labml.ai/

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[41] Y. Ren, J. He, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "Popmag: Pop music accompaniment generation," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1198–1206.

[42] H. Scheffe, *The analysis of variance*. John Wiley & Sons, 1999, vol. 72.